中文

工业级私有 RAG 系统工程叙事

一句话版本 (Executive Summary)

我们正在构建的不是“一个接入大模型的应用”，而是一套可控、可解释、可规模化的私有知识智能系统。该系统通过结构化文档理解、向量召回、二阶段排序和推理型大模型协作，实现企业级知识的稳定召回与知识生成。可直接用于研发知识库、技术文档问答、合规/法务支持与内部专家系统。

一、业务问题：为什么传统 RAG 不够用

当前多数 RAG 系统存在以下致命问题：

1. **召回不稳定**：相同问题多次查询结果不同，无法作为生产系统。
2. **结果不可解释**：无法说明“答案来自哪里”，不具备审计能力。
3. **长文档难解**：PDF / 技术文档不能被切分 chunk，语义断裂。
4. **幻觉频发**：大模型盲目生成，缺乏证据约束。

英语

Industrial-Level Private RAG System Engineering Narrative

One-Sentence Version (Executive Summary)

We are not building "an application that connects to a large model," but rather a controllable, explainable, and scalable private knowledge intelligence system. This system achieves stable recall and knowledge generation at the enterprise level through structured document understanding, vector recall, two-stage sorting, and collaborative reasoning with large models. It can be directly used for research knowledge bases, technical document Q&A, compliance/legal support, and internal expert systems.

I. Business Problem: Why Traditional RAG Is Not Enough

The current majority of RAG systems have the following fatal problems:

1. **Unstable Recall**: Different query results for the same question multiple times, cannot be used as a production system.
2. **Unexplainable Results**: Cannot explain "where the answer comes from", lacks auditability.
3. **Difficult to Handle Long Documents**: PDF/technical documents cannot be split into chunks, semantic fragmentation.
4. **Frequent Hallucinations**: Large models generate blindly, lack evidence constraints.

日语
工業レベルのプライベート RAG システム工学ナラティブ

一文バージョン（エグゼクティブサマリー）
私たちは「大規模モデルに接続するアプリケーション」ではなく、制御可能で説明可能かつスケーラブルなプライベート知識インテリジェンスシステムを構築しています。このシステムは、構造化ドキュメント理解、ベクトルリコール、二段階ソート、および推論型大

規模モデルとの協調を通じて、エンタープライズレベルの知識の安定したリコールと知識生成を実現します。研究知識ベース、技術文書 Q&A、コンプライアンス/法務サポート、内部専門家システムに直接使用できます。

---

1. 業務上の問題：なぜ従来の RAG が不十分なのか
現在の多くの RAG システムには以下の致命的な問題があります：

1. **リコールの不安定性**：同じ質問を複数回問い合わせると結果が異なるため、生産システムとして使用できません。
2. **結果の説明不可能性**：「答えがどこから来たのか」を説明できないため、監査機能がありません。
3. **長文書の処理困難さ**：PDF/技術文書をチャンクに分割できず、意味が断片化します。
4. **頻繁な幻覚発生**：大規模モデルが盲目的に生成し、証拠の制約がありません。

这些问题决定了：

- 普通 RAG 只能用于 Demo，无法用于核心业务。

二、我们的解决方案：工程化的 RAG Pipeline

本系统采用工业级分层架构，每一层职责清晰，可独立替换。

1. 结构化文档理解 (PageIndex)

- **对 PDF / 技术文档进行章节级、语义级组织建模**
- 每个知识单元包含：
  - 文档标题
  - 章节路径 (Section Path)
  - 原文 + 摘要
- **价值**：
  - 避免语义被粗暴切断，支持“章节级精准引用”。

2. 向量召回 (logse-m3)

- **对结构化文本生成 embedding**

- 查询时进行 Top-K 初筛

These issues determine:
- Ordinary RAG can only be used for Demo, not for core business.

II. Our Solution: Engineered RAG Pipeline

This system adopts an industrial-grade layered architecture, with each layer having clear responsibilities and being independently replaceable.

1. Structured Document Understanding (PageIndex)
- **Organize PDF/technical documents at the chapter level and semantic level**
- Each knowledge unit includes:
  - Document title
  - Chapter path (Section Path)

  - Original text + summary
- **Value**:
  - Avoids semantic meaning from being rudely cut off, supports "precise citation at the chapter level".

2. Vector Recall (logse-m3)
- **Generate embeddings for structured text**
- Perform Top-K preliminary screening during queries.

これらの問題は次のように決定します：
- 普通の RAG はデモにしか使用できず、コアビジネスには使用できません。

---

2. 私たちの解決策：エンジニアリングされた RAG パイプライン

このシステムは、各層が明確な責任を持ち、独立して置き換え可能な産業レベルの階層構造を採用しています。

1. 構造化ドキュメント理解（PageIndex）

- **PDF/技術文書を章レベルと意味レベルで組織化する**
- 各知識ユニットには次のものが含まれます：
  - ドキュメントタイトル
  - 章パス（セクションパス）
  - 元のテキスト+要約
- **価値**：
  - 意味が乱暴に切られることを避ける、"章レベルの正確な引用"をサポートする。

2. ベクトルリコール（logse-m3）

- **構造化テキストに対して埋め込みを生成する**
- クエリ時に Top-K の初期スクリーニングを行う。

中文：

- **定位**：

- 负责“语义相关性”，追求高 Recall，而非最终排序。

总结

通过结构化文档理解、向量召回、二阶段排序和推理型大模型协作，本系统解决了传统 RAG 的召回不稳定、不可解释、长文档难解和幻觉频发等问题，为企业级知识管理提供了可靠、高效、可扩展的解决方案。

English:

- **Positioning**:

- Responsible for "semantic relevance", pursuing high Recall, rather than the final ranking.

Summary

Through structured document understanding, vector recall, two-stage sorting, and collaborative reasoning large models,

this system solves the problems of unstable recall, unexplainability, difficulty in understanding long documents, and frequent hallucinations in traditional RAG, providing reliable, efficient, and scalable solutions for enterprise-level knowledge management.

日本語:
- **位置付け**:
- 「意味の関連性」を担当し、最終的なランキングではなく、高いリコールを追求します。

まとめ
構造化されたドキュメント理解、ベクトルリコール、二段階ソート、および推論型大規模モデルの協調により、このシステムは伝統的な RAG のリコールの不安定さ、説明不能性、長文書の難解さ、および頻繁な幻覚などの問題を解決し、企業レベルの知識管理に信頼性が高く、効率的で拡張可能なソリューションを提供します。

以下是图像中识别到的所有文字：

中文
以下是图像中识别到的所有文字：

Query Rewrite (DeepSeek V3)
- 将用户短查询重写为清晰、语义完整的检索意图

示例：
```
plaintext
model train → Transformer 模型的训练流程、训练策略与训练设置
```

价值：
- 显著提升内召回与 reranker 效果 - 属于高 ROI 的“检索增强链”

英文
Query Rewrite (DeepSeek V3)
- Rewrite user short queries into clear and semantically complete search intentions

Example:
```
plaintext
model train → Training process, training strategy, and training settings for the Transformer model
```

Value:
- Significantly improve internal recall and reranker performance - belongs to the high ROI "search enhancement chain"

日語

クエリの書き換え（DeepSeek V3）

- ユーザーの短いクエリを明確で意味的に完全な検索意図に書き換える

例：

```
プレーンテキスト
モデルの訓練 → Transformer モデルの訓練プロセス、訓練戦略、および訓練設定
```

価値：

- 内部リコールとリランカーの効果を大幅に向上させる - 高 ROI の「検索強化チェーン」に属する

工程实践：

- 可与向量分数加权融合 - 可引入 Section / 标题权重

最终生成（DeepSeek R1）

- 不直接“凭空回答”
- 只基于已召回、已排序的证据
- 负责：
  - 归纳
  - 总结
- 结构化表达

定位：

R1 是“答案编辑器”，不是“知识来源”。

三、系统优势（对管理层最重要）

可解释

Engineering Practice:
- Can be integrated with weighted vector scores - can introduce Section / title weight

---

Final Generation (DeepSeek R1)
- Does not directly "answer out of thin air"
- Only based on recalled and sorted evidence
- Responsible:
  - Summarize
  - Conclude
- Structured expression

Positioning:
R1 is an "answer editor," not a "source of knowledge."

Three, System Advantages (Most Important for Management)

Explainable

工学実践：
- ベクトルスコアの重み付けと統合可能 - セクション/タイトルの重みを導入可能

最終生成（DeepSeek R1）
- 直接「空から回答」しない
- 呼び出され、並べ替えられた証拠に基づくのみ
- 担当：
  - 要約
  - 総括
- 構造化表現

位置づけ：

R1 は「答えの編集者」であり、「知識の源」ではない。

三、システムの優位性（管理層にとって最も重要）

解釈可能

中文：

- 每条回答可追溯到具体文档与章节

可控

- 模型只在“被允许的证据范围内”生成

可扩展

- embedding / reranker / LLM 均可替换

私有化

- 全链路支持内网部署
- 数据不出域

希望这些信息对你有帮助！如果有其他问题，请随时告知。

English:

- Each answer can be traced back to specific documents and chapters

Controllable

- The model generates only within the "allowed evidence range"

Scalable

- Embedding / reranker / LLM can all be replaced

Private

- Full chain supports intranet deployment
- Data does not leave the domain

I hope this information is helpful to you! If you have any other questions, please let me know.

日本語：

- 各回答は具体的な文書と章に遡ることができます

制御可能
- モデルは「許可された証拠範囲内」でのみ生成されます

拡張可能
- embedding / reranker / LLM はすべて置き換え可能です

プライベート化
- 全チェーンが社内ネットワーク展開をサポートします
- データはドメイン外に出ません

これらの情報があなたに役立つことを願っています！他の質問があれば、いつでもお知らせください。

以下是图片中识别到的所有文字：

四、应用场景

- 企业内部技术知识库
- 研发 / 运维 / 架构文档问答
- 法务 / 合规文档检索
- 私有 AI Copilot

五、商业价值判断

在真实企业场景中：
- 同类系统的外包成本：300~1000 万人民币
- SaaS 年费：50~200 万 / 年
- 且多数不支持私有化与深度定制

而本系统：

- 架构已成型

Four, Application Scenarios

- Internal technical knowledge base of the enterprise
- R&D / operation and maintenance / architecture document Q&A
- Legal / compliance document search
- Private AI Copilot

Five, Commercial Value Judgment

In real enterprise scenarios:

- Outsourcing cost of similar systems: 300~1000 million RMB
- SaaS annual fee: 50~200 million / year

- And most do not support private deployment and deep customization

For this system:
- The architecture has been formed

四、アプリケーションシナリオ

- 企業内の技術知識ベース
- 研究開発 / 運用保守 / アーキテクチャドキュメントの質問応答
- 法務 / コンプライアンスドキュメントの検索
- プライベート AI コパイロット

五、商業価値判断

実際の企業シナリオにおいて：

- 同様なシステムの外注コスト：3 億～10 億円
- SaaS 年間料金：5 千万～2 億円/年
- かつ大多数がプライベート化と深層カスタマイズをサポートしていない

一方、本システム：
- アーキテクチャは既に形成されています

中文：

- 核心难点已验证
- 具备复用与复制能力

六、结论

这不是一个“玩大模型的项目”，而是：

- 一套可长期演进的企业级知识智能基础设施。

其真正价值不在于模型参数，而在于：

- 工程结构
- 系统边界
- 可解释性与稳定性

这是少数真正能进入生产环境的 AI 系统形态。

希望这些信息对您有所帮助！

English:

- Core challenges have been verified
- Has the ability to reuse and replicate

Six, Conclusion

This is not a "play big model project", but rather:

- A set of enterprise-level knowledge intelligence infrastructure that can evolve over the long term.

Its true value does not lie in the model parameters, but in:

- Engineering structure
- System boundaries
- Explainability and stability

This is one of the few AI system forms that can truly enter the

production environment.

Hope this information is helpful to you!

日本語:
- 核心的な課題はすでに検証済みです
- 再利用と複製の能力を備えています

六、結論

これは「大規模モデルを遊ぶプロジェクト」ではなく、以下のようなものです：
- 長期的に進化できる企業レベルの知識インテリジェンス基盤。

その真の価値はモデルパラメータにではなく、以下の点にあります：

- エンジニアリング構造
- システム境界
- 解釈可能性と安定性

これは、生産環境に入ることができる数少ない AI システム形態です。

これらの情報があなたに役立つことを願っています！